MONTCAS, PHASE 2
Criterion-Referenced Test Alternate Assessment (CRT-Alternate)

2007
TECHNICAL MANUAL

Linda McCulloch, Superintendent

Montana Office of Public Instruction
PO Box 202501
Helena, Montana 59620-2501
www.opi.state.mt.us

TABLE OF CONTENTS

SECTION I: ASSESSMENT DEVELOPMENT

CHAPTER 1—BACKGROUND AND OVERVIEW

1.1 Purpose of This Manual

The purpose of this manual is to document the technical aspects of the 2007 MontCAS, Phase 2 Criterion-Referenced Test Alternate Assessment (CRT-Alternate). In the spring of 2007, students in grades 3 through 8 and 10 participated in the administration of the CRT-Alternate; during this administration, reading and mathematics were assessed. This represents the second year of the CRT-Alternate program, which will expand next year to include science in grades 4, 8, and 10. This report provides information about the technical quality of those assessments, including a description of the processes used to develop, administer, and score the tests and to analyze the test results.

Historically, while some parts of a technical report may have been used by educated laypersons, the intended audience were experts in psychometrics and educational research. This edition of the CRT-Alternate technical report is an attempt to make the information contained herein more accessible to educated laypeople by providing richer descriptions of general categories of information. In making some of the information more accessible, we have purposefully preserved the depth of technical information that has historically been provided in our technical manuals. The reader will find that some of the discussion and tables continue to require a working knowledge of measurement concepts such as "reliability" and "validity," and statistical concepts such as "correlation" and "central tendency." To fully understand some data, the reader will also have to possess basic familiarity with advanced topics in measurement and statistics.

1.2 Purpose of the CRT-Alternate

The Individuals with Disabilities Education Act (IDEA) requires that students with disabilities be included in each state's system of accountability and that students with disabilities have access to the general curriculum. The No Child Left Behind Act (NCLB) also speaks to the inclusion of all children in a state's accountability system by requiring states to report student achievement for all students, as well as for groups of students on a disaggregated basis. These federal laws reflect an ongoing concern about equity: all students should be academically challenged and taught to high standards. It is also necessary that all students be involved in the educational accountability system.

To ensure the participation of all students in the state's accountability system, Montana has developed the Criterion-Referenced Test Alternate Assessment (CRT-Alternate). The CRT-Alternate is a performance-based test that is aligned with Montana's Content Standards and Expanded Benchmarks and measures student performance based on alternate achievement standards. It is expected that only those IDEA-eligible students with the most significant cognitive disabilities will participate in the CRT-Alternate.

The CRT-Alternate is based on, and aligned to, Montana's Content Standards and Expanded Benchmarks in reading and mathematics. Montana educators worked with OPI and its contractor, Measured Progress, in the development and review (content and bias) of these tests to assess how well students have learned the Montana Content Standards and Expanded Benchmarks for their grade. The United States Department of Education (USDOE) approved the CRT-Alternate in reading and mathematics for grades 3–8 and 10 by school year 2005–2006 and in science at one grade in each of three grade spans (e.g., 4, 8, and 10) by school year 2007–2008.

The underlying principal of the assessment is that all students should be taught using Montana's Content Standards and Expanded Benchmarks in reading and mathematics. The tests are intended to measure how a student is performing in relation to those content standards. Results should be used to inform future instruction in the Montana Content Standards.

This was the fourth year of implementation. After the first year, extensive revisions were made based on feedback from teachers who administered the assessment. Alternate assessments have only been in place since 2000. The field is still in the learning stages as to appropriate ways to address reliability and validity for alternate assessments.

To address reliability, several analyses were conducted, including:

- reliability of performance-level categorization
- accuracy
- consistency
- calculating accuracy
- calculating consistency
- kappa

The summary for these analyses can be found in chapter 10. Each chapter in this manual contributes important information to the validity argument by addressing one or more of the following aspects of the CRT-Alternate:

- test development
- test alignment
- test administration
- scoring item analyses
- reliability
- scaling
- performance-levels
- reporting

These aspects, as well as other information on validity, can be found in chapter 13.

1.3 Test Scheduling

The CRT-Alternate was given during the spring: **reading** and **mathematics** were administered in grades 3—8 and 10 during a six-week window (February 12–March 28, 2007). Schools were able to schedule testing sessions at any time during this period. This window, longer than that for the CRT, allowed teachers administering the CRT-Alternate extra time to prepare and adapt test activity materials needed for testing.

The CRT-Alternate is an untimed assessment. Teachers administering the assessments were instructed to watch students for indications that a break may be needed. Recommendation for breaks were inserted in the test booklet. Teachers could choose to stop at the breaks inserted or at other points in the assessment.

1.4 Organization of This Manual

The organization of this manual is based on the conceptual flow of an assessment's life span. It begins with the initial test specification and addresses all the intermediate steps that lead to final score reporting.

Section I covers the development of the CRT-Alternate tests. It consists of six chapters covering:

- Background and Overview
- Inclusion
- Overview of Test Design
- Test Development Process
- Design of the Reading Assessment
- Design of the Mathematics Assessment

Section II consists of a single chapter:

- Test Administration

Section III consists of six chapters covering:

- Scoring
- Item analysis
- Reliability
- Scaling
- Reporting
- Validity

Section IV contains references, and Section V contains the appendices.

CHAPTER 2—INCLUSION

2.1 Sample Size

Because the general CRT provides full access to the vast majority of students, it is expected that only approximately 100 students per grade will participate in the CRT-Alternate. Due to very small sample sizes (70 to 105 students in each grade/content combination), it is unreasonable to calculate Differential Item Functioning (DIF) statistics for the Montana CRT-Alternate. That is, Type I error rates would be unreasonably high and would result in incorrect conclusions regarding the functioning of the items between reference and focal groups. Thus, DIF statistics are not included as part of this technical report.

Number of Students Participating in Each Assessment for Spring 2007		
Grade	**Content Area**	**N**
3	Mathematics	68
	Reading	69
4	Mathematics	90
	Reading	90
5	Mathematics	73
	Reading	71
6	Mathematics	108
	Reading	107
7	Mathematics	73
	Reading	71
8	Mathematics	72
	Reading	72
10	Mathematics	107
	Reading	107

In accordance with 34 CFR 200.13 Adequate Yearly Progress in general, there is a 1% cap applied to the number of proficient and advanced scores based on the alternate assessment that may be included in AYP calculations at both the state and district levels.

2.2 Participation Guidelines

The decision as to how a student with disabilities will participate in the state's accountability system is made by the student's Individualized Education Program (IEP) team. When considering

whether students with disabilities should participate in the CRT-Alternate, the IEP team should address each of the questions in the chart that follows:

Participation Guidelines:		
For each of the statements below, answer YES or NO.		
Does the student have an active IEP and receive services under the Individuals with Disabilities Education Act (IDEA)?	YES	NO
Do the student's demonstrated cognitive abilities and adaptive behavior require substantial adjustments to the general curriculum?	YES	NO
Do the student's learning objectives and expected outcomes focus on functional application of skills, as illustrated in the student's IEP's annual goals and short-term objectives?	YES	NO
Does the student require direct and extensive instruction to acquire, maintain, generalize and transfer new skills?	YES	NO

- **If you answer "NO" to <u>any</u> of the above questions, the student must participate in the regular CRT.**
- **If all answers are "YES," the student is eligible to take the alternate assessment and is considered to be a student with a significant cognitive disability.**

The decision to determine a student's eligibility to participate in the CRT-Alternate <u>may not be based on</u>

- excessive or extended absence;
- disability category;
- social, cultural, or economic difference;
- the amount of time receiving special education services; or
- academic achievement significantly lower than his or her same-age peers.

CHAPTER 3—OVERVIEW OF TEST DESIGN

3.1 CRT-Alternate

CRT-Alternate test items are directly linked to **Montana's Content Standards and Expanded Benchmarks**. (See page 19 for more information about the expanded benchmarks.) The content standards are the basis for the reporting categories developed for each content area and are used to help guide the development of test items. No other content or process is subject to statewide assessment. An item may address part, all, or several of the benchmarks within a standard or standards.

3.2 Assessment Types

Although the CRT-Alternate for all grades is a performance-task assessment, the format differs slightly depending on the grade assessed. This difference is due to the fact that the assessments for grades 4, 8, and 10 were developed two years earlier than the assessments for grades 3, 5, 6, and 7. All assessments follow the same scaffolding rubric, follow the same expanded benchmarks, and are designed to show a student's performance in relation to the Montana reading and mathematics standards and benchmarks. However, there are some notable differences between the two formats. The major differences are outlined below:

Highlighted Differences Between the Two Assessment Formats		
Topic	**CRT-Alternate for Grades 3, 5, 6, and 7**	**CRT-Alternate for Grades 4, 8, and 10**
Format	• Tasklet—five short activities of five items each per content • Total of 25 items	• One overall activity with 22–35 items per content
Introductory Items	• First item in each tasklet designed to get student's attention, introduce the activity, and show materials to be used • Scored at levels 4 or 0 of the rubric	• First few items in each activity and may have one or more interspersed as new materials are introduced in later sections of the activity • Designed to get student's attention, introduce the activity, and show materials to be used • Scored at levels 4 or 0 of the rubric
Breaks	• Breaks between tasklets	• Suggested breaks built into activity
Reading Passage	• Page 2 of each reading tasklet	• Grade 4 only: page 2 of the reading activity
Student Evidence	• 1–2 tasklets in each content require student evidence	• Each overall activity requires evidence • Two forms need to be filled out for

	• Two forms need to be filled out for each item that requires evidence	each item that requires evidence
Scoring Rule	• Student must try every tasklet • Halt the administration of a tasklet only if the student scores a 0 for three consecutive items after administering the tasklet in two different test sessions	• Halt the administration of the activity after the student scores a 0 for three consecutive items after administering the activity in two different test sessions
Materials Kits	• Tabs in the Materials Kits are labeled by content and tasklet number	• Tabs in the Materials Kits are labeled by content and separated by Activity Materials (A.M.) and Communication Supports (C.S.). Within the two sections, tabs are labeled evidence templates, sentence strips, four-choice grids, number cards, etc.

After completing the assessment, the teacher was asked to respond to a series of questions regarding preparation and administration. Question 11 asked the teacher to report how much time he or she spent in preparing for the assessment. Question 12 asked the teacher to report how much time was spent administering the assessment. According to the embedded teacher survey, the activities in grades 4 and 8 seemed to take an average amount of time to prepare for, but generally less time to administer than the tasklets at grades 3, 5, 6, and 7. Grade 10 was reported as having the lowest average preparation and administration times of any grade. A summary of responses to questions 11 and 12 are summarized in the tables below.

Survey Response Question 11—Set Up Time/Planning		
Grade	Subject	Average # of Hours
3	Reading	1.02
3	Math	0.95
4	Reading	1.05
4	Math	1.03
5	Reading	0.99
5	Math	0.95
6	Reading	1.31
6	Math	1.27
7	Reading	1.22
7	Math	1.32
8	Reading	1.47
8	Math	1.48
10	Reading	0.95
10	Math	0.82

Survey Response Question 12—Time directly administering the assessment		
Grade	Subject	Average # of Hours
3	Reading	1.41
3	Math	1.39
4	Reading	1.14
4	Math	1.18
5	Reading	1.31
5	Math	1.22
6	Reading	1.36
6	Math	1.29
7	Reading	1.46
7	Math	1.47
8	Reading	1.11
8	Math	1.31
10	Reading	1.01
10	Math	1.06

Assessment Type for Grades 4, 8, and 10

The CRT-Alternate assessment is a point-in-time test that looks at how students perform in relation to performance indicators that have been expanded from the Montana reading and mathematics standards and benchmarks. Each content area in grades 4, 8, and 10 consists of one age-appropriate activity that has 20 to 35 items in which the teachers are given a script, written directions, and scaffolding levels. Students are encouraged to engage in the activity and show performance on the items through appropriate prompting by the teacher who administers the test activity. The teacher who administers the test activity scores the student on each item through observation using a five-point scoring rubric.

The test activity requires evidence to be collected based on the products that were created during the course of the assessment. Templates were provided for all evidence that was required.

Assessment Type for Grades 3, 5, 6, and 7

The CRT-Alternate assessment is a point-in-time test that looks at how students perform in relation to performance indicators that have been expanded from the Montana reading and mathematics standards and benchmarks. Each content area in grades 3, 5, 6, and 7 have five tasklets (short activities) that consist of five questions each in which the teachers are given a script, written directions, and scaffolding levels. This tasklet format allows for natural breaks in the assessment, so the student may rest and refocus between tasks. Students are encouraged to engage in the activity and

show performance on the items through appropriate prompting by the teacher who administers the test activity. The teacher who administers the test activity scores the student on each item through observation using a five-point scoring rubric.

The test activity requires evidence to be collected based on the products that were created during the course of the assessment. Templates were provided for all evidence that was required.

3.3 CRT-Alternate Items (all grades)

Each item of the CRT-Alternate consists of the following:

- materials needed to administer the item
- setup instructions and script for the teacher to follow if using the test activity
- scaffolding script for the suggested test activity
- the correct student response
- the performance indicator (The performance indicator is what the question is measuring. The performance indicator comes from the Montana Content Standards and Expanded Benchmarks.)
- activity steps to follow for teachers creating their own activity

See chapter 2 for the test format.

3.4 Scaffolding as Scoring

As Gail McGregor of the University of Montana—Missoula notes in her paper titled "Implementation of the CRT-Alternative Strategies to Achieve Interrater Reliability" (Appendix H), "Administration of the CRT-Alt incorporates a response prompting methodology known as the 'system of least prompts' (Wolery, Ault & Doyle, 1992), a well-established strategy that has been found to be effective as a teaching procedure for students with severe disabilities across a wide range of applications (Doyle, Wolery, Ault & Gast, 1988)." The system of least prompts, or scaffolding, requires the teacher (or test administrator) to administer each test item beginning at the highest level of independence. The student is asked the question and allowed sufficient time to produce the answer. If the student produces the answer, the teacher records his or her score for that question at the highest level. If the student answers incorrectly, the test administrator asks the question again but this time using the next-to-highest level of independence for this particular question. The levels of independence are standardized and scripted within the test. This second-highest level of independence usually amounts to removing one or two choices from the set of posssible answers. If the student provides the

correct answer this time, the test administrator will record the score at this second-highest level of independence. If the student cannot provide the correct answer, the test administrator moves on to the next-highest level of independence, and so on, until the student is guided (hand-over-hand) to the correct answer and the student's score for that particular item is recorded at the lowest level of independence. More information regarding the research base of this method and a discussion regarding the selection of this method can be found in appendix H.

3.5 Test Format

Grades 4, 8, and 10

In grades 4, 8, and 10, the CRT-Alternate is composed of two test activities: one for reading and one for mathematics. Each test activity consists of 20 to 35 items and at least one piece of student evidence (work). Since only one test was developed, every student takes the same form of the test. The test stays the same each year, with the exception of the second year, when revisions were made using teacher feedback during a revision workshop. Although the test items are kept secure, the performance indicators, which come from the Montana reading and mathematics Content Standards and Expanded Benchmarks, are released every year on the OPI and Measured Progress Web sites. The 2007 released performance indicators are located in appendix E.

Grades 3, 5, 6, and 7

In grades 3, 5, 6, and 7, the CRT-Alternate is composed of ten tasklets: five for reading and five for mathematics. Each tasklet consists of five items relating to the small activity. Some tasklets require student evidence, and some do not. Creating the test around a series of smaller activities (rather than one single activity as in grades 4, 8, and 10) allows the teacher and student to break the administration into smaller time segments without being as concerned about a disruption in continuity. Since only one test was developed, every student takes the same form of the test. Although the test items are kept secure, the performance indicators, which come from the Montana reading and mathematics Content Standards and Expanded Benchmarks, are released every year on the OPI and Measured Progress Web sites. The 2007 released performance indicators are located in appendix E.

The first page of each activity (in grades 4, 8, and 10) or tasklet (in grades 3, 5, 6, and 7) lists the following:

- content standards

- a brief explanation of the suggested test activity
- parameters of the task
- materials provided and other materials that are needed

The pages that follow in the math and reading sections of the test booklet consist of the following four columns for each item:

Materials for the Activity	Activity Teacher will:	Student Work Student will:	Performance Indicators Use Scoring Guide Transfer scores to student response booklet
The materials that are needed for each item and suggested student communication supports and strategies that may be helpful for some students are described in this column. Most materials can be found in the Materials Kits, but teachers need to supply some materials.	**This column contains information about how to display task materials and prepare the student for the question. A script for the teacher appears in bold and italicized print and suggests language that can be used to present the item.** **Information on how to scaffold levels 3, 2, and 1 of the rubric for items that are scored at levels 4 through 0 is also provided in this column.**	**The correct student response and/or an explanation of how the student should be responding is provided in this column.**	**The performance indicator that is assessed by each item is identified in this column. The performance indicators come from the Montana Content Standards and Expanded Benchmarks.**
SAMPLE TASKLET (GRADE 4 MATH)			
2. 1 square, 1 circle, 1 rectangle, and 1 triangle. **Communication support strategies:** • **Student may look at/point to task materials to express a choice.** • **Request may be rephrased to require a yes/no response (e.g., "Is this the square?").**	**2. Place the shapes on the work space.** **"Here are the shapes we just looked at. Show me the square."** **Scaffold:** **Level 3: Remove an incorrect response. Repeat task request.** **Level 2: Remove another incorrect response. Repeat task request.**	**2. Indicate the square.**	**2. Identify (name) shapes as circles, squares, triangles, rectangles, and ovals.** o o o o o 4 3 2 1 0 **Performance Indicator:** **4.1.1.6** **Expanded Benchmark:** **4.1.1**

• **Student may tell teacher to "stop" at desired response as teacher sequentially points to each of the 4 choices.**	**Level 1: "This is the square." Assist the student as needed to identify the square.**

Evidence and Evidence Template(s)

Each of the test activities requires that evidence be collected based on the products that are created during the course of the assessment. A magnifying glass in the "Student Work, Student will" column of the test booklet indicates when evidence must be collected. Templates are provided in the CRT-Alternate test booklet for all evidence that is required. Teachers have the option of selecting the presentation that best matches the student's abilities and skills:

- written work by the student (e.g., the student collects data and fills out a bar chart with a marker)
- pictures of student output (e.g., the student arranges objects to form an answer to a question about the sequence of events in a story, and a picture captures the arrangement)
- picture symbols pasted on the template or a scanned/photocopied image of the template that the student arranges and that he or she wants to keep
- computer printout of the student's keyed responses
- teacher-recorded responses (e.g., the teacher fills out a T-table based on the yes/no answers from the student using a BIGmack switch or eye gaze)
- anecdotal record describing the student's actions supplied by the observer (e.g., the observer notes that the student smiled when shown a picture of his or her favorite character in a story)

The evidence templates are used to record student responses to an item when asked. Adapted evidence templates are provided in the Materials Kits and on the Materials CD. The template may need further modifications based on the student's needs. The evidence must be submitted along with the used test booklet. Upon receipt, evidence is scanned and accounted for. OPI was provided with a list of students (and their schools) who did not provide evidence along with their test booklets.

Last Page of the Test Booklet

The last page of the test booklet contains a list of questions for the teacher to answer after the administration of the reading and mathematics test activities.

CHAPTER 4—TEST DEVELOPMENT PROCESS

4.1 Item and Activity Development

The CRT-Alternate was developed as a collaborative project between Measured Progress and the Montana Office of Public Instruction (OPI) divisions of Assessment, Special Education, and Educational Opportunity and Equity.

An advisory committee, representing perspectives of parents, teachers, administrators, and faculty in higher education, provided input during the development of this assessment. In addition, teacher work groups were formed at several points in the development and revision process. Mathematics and reading item development work groups were composed of general and special education teachers. These teachers developed test activities that are the basis of the performance tasks for this assessment. A third group of special education teachers and administrators participated in the beta testing of this assessment, providing valuable feedback about the test design.

OPI was responsible for organizing and facilitating committees to review items and reading passages for bias and sensitivity. OPI sent the feedback from the committees to Measured Progress to make the appropriate changes to the items and reading passages.

4.2 Development of the Reading and Mathematics Expanded Benchmarks

The expanded benchmarks were developed for students with significant cognitive disabilities not working at the same level as their age-level counterparts. The expanded benchmarks were developed using Montana's Content Standards and Expanded Benchmarks for reading and mathematics. Measured Progress's curriculum and special education specialists developed a draft of the expanded benchmarks. The OPI, beta test teachers, the advisory committee, and the development and revision workshop participants all provided input and recommendations for changes to the original draft. Using these recommendations, Measured Progress revised the expanded benchmarks. This document was further revised to include grade-span expectations per new federal legislation. It is expanded from end of grade 4, end of grade 8, and end of grade 12—upon graduation to foundational skills. These are grade-span expectations due to the wide diversity of students in this population. This document was used to develop the assessment performance indicators. The chart on the next page shows how the document is organized and gives an example for each content area. The Montana

Content Standards and Expanded Benchmarks are not included in this manual because of the length of each document. They are located on the OPI Web site at www.opi.state.mt.us and the Measured Progress Web site at www.measuredprogress.org.

Montana CRT-Alternate Standards and Expanded Benchmarks

Terminology		
Term/Description	**Example**	
Content Area	**Mathematics**	**Reading**
Standard Learning outcome expected for all students throughout all grades	Standard 2: Students demonstrate understanding of and ability to use Numbers and Operations.	Standard 2: Students apply a range of skills and strategies to read.
Essence of the Standard A statement of the standard separating the essential components	Number concepts, concepts of operations, computing and estimating	Interpret print and nonprint information
Benchmark **Grade Level Expectation (GLE)** Expectation for typical students described for each grade level	2.2, Grade 4: Students will use the number system by counting, grouping, and applying place value concepts.	2.6, Grade 8: Students will develop vocabulary through the use of context clues, analysis of word parts, auditory clues, and reference sources (e.g., dictionary, thesaurus, and glossary).
Expanded Benchmark Benchmark skill or concept expanded from the typical GLE to a basic level	2.2.1: Student will demonstrate an understanding of whole numbers.	2.6.2: Student will use words/pictures/symbols/objects to communicate.
Performance Indicator Expanded benchmark expressed in a measurable and observable statement of a specific performance	2.2.1.2: Student will demonstrate the concept of one (e.g., "Hit the switch one time"; "Give me one").	2.6.2.1: Student will identify a word/picture/symbol/object used to name a familiar place.
Prompt The script for the directions the test administrator will deliver to the student, calling for the specific behavior	Item 4: "These are counters. We are going to use these in our activity. Show me one counter."	Item 4: "Show me the word/picture/symbol/object that means 'library.'"

TOTAL NUMBERS OF ITEMS DEVELOPED BY GRADE AND CONTENT		
GRADE	READING	MATH
3	25	25
4	22	28
5	25	25
6	25	25
7	25	25
8	24	32
10	27	31

4.3 CRT-Alternate Item Development Process Overview

As previously noted, there were separate development process cycles used to create the body of tests that now compose the current CRT-Alternate. Grades 4, 8, and 10 were developed between August 2003 and October 2004. An overview of the test development process for the CRT-Alternate program in grades 4, 8, and 10 is outlined in the technical manual for 2005. The second cycle of development, for alternate assessments in grades 3, 5, 6, and 7, took place between March 2005 and January 2006 and is outlined in the technical manual for 2006. For all grades, the test-development process began with the expansion of benchmarks for reading and mathematics in 2003. Using the expanded benchmarks for reading and mathematics, staff from Measured Progress created a test blueprint for each grade. The blueprint indicated which expanded benchmarks should be tested at each grade. Once the blueprint was approved by the state, development workshops were held. At these development workshops, Montana educators came up with tasklet ideas to use in the creation of the tests. Staff from Measured Progress selected passages for reading and topics for math and began creating draft tasklets. The state was involved at every step in the process to provide feedback for changes to the tasklets or give approval. After the editorial-and-approval phase, the tasklets were beta tested using Montana educators and their students. After the beta test, revisions were made based on feedback from the field.

DEVELOPMENT PROCESS OVERVIEW	
DEVELOPMENT STEP	PROCEDURE OF THE STEP
Development and revision of expanded benchmarks for reading and mathematics	• Measured Progress curriculum and special education specialists developed a draft of the expanded benchmarks. • The OPI reviewed it. • Beta test teachers provided input.

	• The advisory committee and revision and development workshop participants provided recommendations. • The expanded benchmarks were revised to include grade-span expectations per new federal legislation.
Blueprint design	• Measured Progress curriculum and special education specialists created initial assessment blueprint. • Blueprint was approved by the state.
Development workshops	Measured Progress curriculum and special education specialists and the OPI: • provided item development training to Montana participants; • facilitated the development of the item ideas by the participants.
Passage/topic selection and development	Reading passages and mathematics topics were selected for the tasklets: • Measured Progress used the items and activities that were developed at the development workshops to prepare topics and passages for the state; • The state was given the topics and passages to approve; • The state made approvals.
Tasklet creation	Measured Progress curriculum and special education specialists: • used the blueprint, tasklet ideas, and passages/topics to create test items (tasklets).
Editorial review of items	All items were reviewed by members of the Measured Progress publications staff to ensure: • clarity and unambiguousness of items; • correct grammar, punctuation, usage, and spelling; • technical quality with respect to stems, options, and scoring guides; • compliance with OPI sensitivity standards and style guidelines.
Beta test	• Approximately 20 students participated in the beta test. • Beta test teachers tested a student on one content area and sent feedback to Measured Progress on the assessment items and activity. • Beta test participants gave additional feedback in a conference call. • The Advisory Committee reviewed all grades and contents and provided feedback via a form and conference call.
Revisions after beta test	• Using the feedback from the beta test teachers and the advisory committee, the OPI and Measured Progress revised the assessment. • Level 1 scaffolding script was added to every item on the test that is scored using all five levels of the rubric.

4.4 Revisions Made to the Spring 2005 Assessments for Grades 4, 8, and 10

Using feedback from teachers who administered the CRT-Alternate in the spring of 2004, Montana special education and general education teachers, the OPI, and Measured Progress revised the following in the assessments:

- Level 1 scaffolding language was added to the "Activity, Teacher will" column. This was added to give teachers a clearer direction on how to scaffold this level.
- The "Materials for the Activity" column was added. This column lists the materials needed for each item, as well as communication-support strategies. This column was added to prepare teachers on what materials are needed to administer each item and for students to respond to each item. It also gives teachers ideas for student communication supports.
- Ancillary materials and training CDs were developed and sent to teachers administering the assessment.
- Optional breaks were added to give teachers a clearer idea of when to give the student a break in the test activity.
- Item language was revised for clarity and consistency with the newly developed assessments in grades 3, 5, 6, and 7.
- Items were added and deleted to help cover all standards evenly across all grades (3–8 and 10).
- The scoring rule for halting the assessment was changed from "Score every item until the student scores in level 1 or 0 for five consecutive items. Halt the administration if the student scores in level 1 or 0 for five consecutive items. Leave the remaining items blank" to "Score every item until the student scores at level 0 for three consecutive items. Stop the administration of the assessment at this point. On the following assessment session, readminister the final three items on which the student scored a 0. If the student receives a level 0 on three consecutive items again, halt the administration of the assessment and leave the remaining items blank." Three examples were given for this new rule. This was based on in-depth discussion with the Technical Advisory Committee (TAC) and their recommendations.

4.5 Activities Versus Tasklets:

The earlier tests, in grades 4, 8, and 10 were designed around a single activity. A series of test items were administered using this common activity. When the new tests for grades 3, 5, 6, and 7 were

developed, it was recommended that instead of using one activity with 25 to 35 associated items, a better approach would be to use five smaller activities with five associated items each. This allows for natural break times in the assessment, so that it can be given over a longer period of time. Using five tasklets instead of one activity also helps to minimize the negative impact on a student's score associated with a student who is unusually distracted by the content of a particular tasklet can have. For instance, in the grade 8 mathematics activity, if a student has some sort of negative reaction to cake (maybe he or she is allergic to flour, for example), the fact that all the questions on the test are somehow related to cake may be difficult for him or her. OPI has asked Measured Progress to consider development work that will modify grades 4, 8, and 10 to mirror the tasklet model in grades 3, 5, 6, and 7. Teachers are particularly interested in having this change made to the assessment for consistency as well as for teacher and student involvement.

4.6 Item/Activity Editing

Editors reviewed and edited the items and test activities to ensure uniform style (based on *The Chicago Manual of Style*) and adherence to sound testing principles. These principles included the stipulation that the items and test activities:

- were correct with regard to grammar, punctuation, usage, and spelling;
- were written in a clear, concise style;
- were measuring the performance indicator;
- had materials that were appropriate;
- contained unambiguous explanations for teachers as to what was required of the student;
- were written at a reading level that would allow the student to demonstrate his or her knowledge of the tested subject matter regardless of reading ability;
- exhibited high technical quality regarding psychometric characteristics;
- had appropriate scaffolding script for teachers; and
- were free of potentially insensitive content.

Items should assess only knowledge or skills that are identified as part of the domain being tested and should avoid assessing irrelevant factors. They should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. Further, items must not unfairly disadvantage test takers from particular racial, ethnic, or gender groups.

Both qualitative and quantitative analyses are conducted to ensure that Montana CRT-Alternate items meet these standards.

CHAPTER 5—DESIGN OF THE READING ASSESSMENT

5.1 Reading Blueprint

As indicated earlier, the framework for reading was based on Montana's reading Content Standards and Expanded Benchmarks, which identify five **content standards** that apply specifically to reading and reading comprehension. Those content standards are:

- **Reading Standard 1:** Students construct meaning as they comprehend, interpret, and respond to what they read.
- **Reading Standard 2:** Students apply a range of skills and strategies to read.
- **Reading Standard 3:** Students set goals, monitor, and evaluate their reading progress. (This standard is not measurable in a statewide assessment.)
- **Reading Standard 4:** Students select, read, and respond to print and nonprint material for a variety of purposes.
- **Reading Standard 5:** Students gather, analyze, synthesize, and evaluate information from a variety of sources and communicate their findings in ways appropriate for their purposes and audiences.

The chart below shows the standards measured at each grade level. For a complete list of all reading and mathematics test items (and the correlating standards assessed through each item), see appendix E.

DISTRIBUTION OF READING STANDARDS MEASURED AT EACH GRADE

	STANDARD 1	STANDARD 2	STANDARD 3	STANDARD 4	STANDARD 5
GRADE 3	**13**	**8**	*****	**4**	**0**
GRADE 4	**9**	**9**	*****	**3**	**1**
GRADE 5	**13**	**8**	*****	**4**	**0**
GRADE 6	**13**	**7**	*****	**1**	**4**
GRADE 7	**13**	**7**	*****	**1**	**4**
GRADE 8	**10**	**10**	*****	**2**	**2**
GRADE 10	**13**	**7**	*****	**3**	**4**

*Standard 3 is not measurable in a statewide assessment.

Note: Grade level test blueprints were designed so that the emphasis on concepts in the CRT-Alternate would reflect emphasis on concepts in the general CRT. Standards 1 and 2 for both math and reading are measured at every grade level, and the other standards are measured evenly across grade spans (elementary 3–5, middle 6–8, and high school 10).

CHAPTER 6—DESIGN OF THE MATHEMATICS ASSESSMENT

6.1 Mathematics Blueprint

The mathematics framework was based on Montana's mathematics Content Standards and Expanded Benchmarks, which identify seven **content standards**, as shown below:

- **Mathematics Standard 1:** Problem Solving
- **Mathematics Standard 2:** Numbers and Operations
- **Mathematics Standard 3:** Algebra
- **Mathematics Standard 4:** Geometry
- **Mathematics Standard 5:** Measurement
- **Mathematics Standard 6:** Data Analysis, Statistics, and Probability
- **Mathematics Standard 7:** Patterns, Relations, and Functions

The chart below shows the standards measured at each grade level. For a complete list of all reading and mathematics test items (and the correlating standards assessed through each item), see appendix E.

DISTRIBUTION OF MATH STANDARDS MEASURED AT EACH GRADE

	STANDARD 1	STANDARD 2	STANDARD 3	STANDARD 4	STANDARD 5	STANDARD 6	STANDARD 7
GRADE 3	9	10	0	10	0	0	5
GRADE 4	9	8	0	0	0	13	4
GRADE 5	8	10	5	0	10	0	0
GRADE 6	6	10	0	5	5	0	5
GRADE 7	9	10	10	0	0	5	0
GRADE 8	7	8	4	0	5	11	0
GRADE 10	5	13	7	4	0	0	3

Note: Grade level test blueprints were designed so that the emphasis on concepts in the CRT-Alternate would reflect emphasis on concepts in the general CRT. Standards 1 and 2 for math are measured at every grade level, and the other standards are measured evenly across grade spans (elementary 3–5, middle 6–8, and high school 10).

SECTION II: TEST ADMINISTRATION

CHAPTER 7—TEST ADMINISTRATION

7.1 Responsibility for Administration

The special education teacher or someone who is certified and has worked extensively with the student and is trained in the assessment procedures administers the assessment. The test administrator may find it helpful to ask another person in the school to assist with the administration. Because this is an on-demand performance assessment, the administrator is also the scorer. This becomes a consideration with regards to reliability; the values tend to be inflated due to administrator effects. This is discussed further in chapter 10—Reliability.

These additional persons may include but are not limited to the following:

- parent
- general education teacher
- paraprofessional
- special service provider (speech/language therapist, psychologist, occupational, or physical therapist, etc.)
- school counselor
- principal
- other educational professional

7.2 Procedures

Teachers administering the CRT-Alternate were sent a training CD with an audio PowerPoint presentation to train them on implementing the test. The following are the procedures teachers were given inn preparation for administering the assessment:

- View training CD and participate in question/answer sessions.
- Receive the secure *CRT-Alternate Test Booklet* from the test coordinator.
- Receive hard copy of the test activity materials, CD with test activity materials, and training CD from Gail McGregor at the Rural Institute of Disabilities, University of Montana—

Missoula. Teachers may have needed to further adapt materials to meet the need of the students taking the assessment. Guidelines and examples for adapting materials were given in the "Materials" section of the test booklet and on pages 28 and 29 of the *CRT-Alternate Administration Manual*.

- Download the *CRT-Alternate Administration Manual* and scoring rubric from the OPI or Measured Progress Web site.
- Read the *CRT-Alternate Administration Manual* to become familiar with the administration and scoring directions.
- Read the *CRT-Alternate Test Booklet* to become familiar with the test activity steps and performance indicators.
- Consider how the student will access and respond to the test activity. Determine the adaptations and supports the student will need.
- Check to ensure all materials and resources needed to complete the test activity are available. For example, the grade 8 reading activity asks the student to locate the library and to identify the librarian. The reference or book area in the classroom may be substituted for the library, and someone who helps the student pick a book (i.e., teacher) may be substituted for the librarian.
- Provide the assistive technologies the student needs to access the materials and respond to the test activities.
- Schedule the assessment administration session for a time and place that are optimal for student effort and focus.

7.3 Training

School test coordinators were instructed to read the *Test Coordinator's Manual* prior to testing and be familiar with the instructions given in the *Test Administrator's Manual* and the *CRT-Alternate Administration Manual*. The *Test Coordinator's Manual* and the *CRT-Alternate Administration Manual* provided each school with checklists to help prepare for testing. The checklists outlined tasks to be performed before, during, and after test administration. Along with providing these checklists, the *Test Coordinator's Manual* and the *CRT-Alternate Administration Manual* outlined the nature of the testing material being sent to each school, how to inventory the material, how to track it during administration, and how to return the material once testing was complete. It also contained information

about including or excluding students. The *CRT-Alternate Administration Manual* included a checklist for the test administrators to prepare themselves, their classrooms, and their students for the administration of the test and how to return the assessment. In addition to distributing the *Test Coordinator's Manual* and *CRT-Alternate Administration Manual*, teacher-training CDs were sent to every teacher administering the CRT-Alternate. Training materials and the PowerPoint presentation were posted on the OPI's Web site. Below is a summary of the information presented in the training CD:

- Important Dates
- CRT-Alternate Overview
- Week 1 of Testing
- Eligibility for the CRT-Alternate
- Who Should Administer the CRT-Alternate
- Materials Needed for this Presentation and for Testing
- About the Tests…
- Test Booklet Organization for Grades 3, 5, 6, and 7
- Test Booklet Organization for Grades 4, 8, and 10
- Assessment Format (All Grades)
- Scoring
- Scaffolding
- Dealing with Resistance
- Scoring Rule Grades 3, 5, 6, and 7
- Scoring Rule for Grades 4, 8, and 10
- Introductory Item
- Student Evidence
- Grade-Specific Information for Administering the CRT-Alternate
- Student Response Booklet (SRB)
- Class Identification Sheet
- Student Barcode Labels
- Returning Student Materials
- Test Administration Strategies
- Test Activity Materials Grades 3, 5, 6, and 7

- Test Activity Materials Grades 4, 8, and 10
- Final Administration Hints
- Questions and Answers

To answer any questions that may not have been addressed in the training, teachers, test administrators, and test coordinators were provided with contact information for OPI, Measured Progress, and the University of Montana—Missoula. The contact information was provided on the training CD, in the manual, and on the memo sent out with the test materials.

SECTION III:
DEVELOPMENT AND REPORTING OF SCORES

CHAPTER 8—SCORING

8.1 Scoring the Assessment

Teachers administer the assessment to a student one-on-one or with the help of another administrator. The teacher scores every item as it is administered using the rubric and a process called scaffolding.

8.2 Using Scaffolding to Gather Student Performance Information

Scaffolding is a process of providing the student with the support needed to respond to the questions in the test activity. During daily instruction, many strategies are used frequently to ensure that students experience success. For example, if a student is unable to make a correct choice from a display of four pictures, the teacher reduces the complexity of the test activity by removing one of the choices. Scaffolding serves this same function and is provided so that students will experience success in completing the test activities. An important result of scaffolding is that it helps students demonstrate their knowledge and skills. These skills can be described and measured, resulting in an accurate picture of what students can do.

The scoring system in the CRT-Alternate is built on increasing amounts of scaffolding, provided only when the student does not respond or responds incorrectly. This approach is sometimes described as a "least to most" prompt hierarchy (see chapter 3 for a description of the scaffolding-as-scoring paradigm).

Each test activity begins with items that introduce the subject and materials that will be used in the test activity. These items are scored as either a 4 (student responds accurately and with no assistance) or a 0 (student does not respond or actively resists). Items that are scored at a level 4 or 0 may also be found further into the activity when new materials are being introduced.

After these items are scored, each subsequent item within the test activity is scored on a five-point scale 4–0, with 4 representing a correct, independent response and 1 representing a correct

response that has been completely guided by the teacher. A score of 0 is used when the student does not respond or actively resists participation in the test activity. See the scoring rubric in this section.

The scores from all items, including the introductory items and the subsequent items within the test activity, are added together to produce a raw score (or total score) for the test. The raw score is then scaled and a performance level is assigned for the content area (see chapter 11 for details on scaling).

A script is provided for scaffolding for each of the suggested test activities. It describes the prompts that can be used to scaffold the student to a level 3, level 2, and level 1. It may be used verbatim or modified by the teacher to meet the needs of the student. For each test item, level 1 prompting is full support from the teacher to guide the student to the correct response. Depending on the student and the test item, this may involve physically guiding the student to the correct response or some other form of support that ensures the student responds correctly.

It is critical that the test administrator deliver each item in a way that allows the student the opportunity to score at level 4. That is, assume that the student can respond independently to each item, even if that is not the usual instructional practice. The following are directions given to test administrators in order to standardize scaffolding procedures across the state:

- Follow the guidelines to observe the student demonstrating the performance required and allow adequate wait time for the student to process the information and respond without assistance. Do not repeat the questions multiple times.
- If the student does not respond or responds incorrectly, scaffold the student to level 3—"Student responds accurately when teacher clarifies, highlights important information, or reduces the range of options to three." Again, give the student adequate wait time.
- If the student does not respond or responds incorrectly, scaffold to level 2—"Student responds accurately when teacher provides basic yes/no questions or forced choices between two options."
- If the student still does not respond with the desired behavior, scaffold to level 1—"Student is guided to correct response by teacher (e.g., modeling the correct response or providing full physical assistance)."
- If the student resists participation for an item, the test administrator will indicate a 0—"Student does not respond or actively resists."

Scaffolding is based on the amount of information the student needs to reach the correct response. If the student can respond independently (4), no further information is needed by the student. If the student does not respond accurately or independently, more information is given about the item, and the choices are reduced (3)—see script in the *CRT-Alternate Test Booklet*. This funneling toward the correct response continues as the student needs more assistance—by providing specific information about the item and a forced choice between two options (2)—see script in the CRT-Alternate Test Booklet, and finally, by guiding the student to the correct response (1)—see script in the *CRT-Alternate Test Booklet*. In this way, the student is not expected to "get it" or "not get it," as in most on-demand assessments. The CRT-Alternate considers the level of assistance that students need to demonstrate their knowledge and skills and thus provides more precise information about student performance and achievement. This system is sensitive to small increments of change in student performance, an important consideration in describing the learning outcomes of students with severe disabilities.

This process must be used systematically with each item identified for scoring within the test activity. The intent is to give the student every opportunity to perform independently on each item. Scaffolding examples are given in the *CRT-Alternate Administration Manual*.

The use of different levels of assistance required during administration/scoring will increase item intercorrelations and overall test reliability. The effects of scaffolding and other scoring analysis are further discussed in chapter 10—Reliability.

8.3 Scoring Rubric

Each test activity begins with introductory items. Only rubric levels of 4 and 0 are used to score these introductory items. Items that are scored at a level 4 and 0 may also be found further into the assessment when new materials are being introduced. All five levels of the rubric are used to score remaining items. Teachers administering the assessment are encouraged to have the rubric available as a reference when giving the test. The five levels of the rubric are on the following page.

Montana Alternate Assessment Scoring Guide

Performance (independence and accuracy)

Used to score every item during the structured observation test activity.

4	3	2	1	0
Student responds accurately and with no assistance.	Student responds accurately when teacher clarifies, highlights important information or reduces the range of options to three.	Student responds accurately when teacher provides basic yes/no questions or forced choices between two options.	Student is guided to correct response by teacher (e.g., modeling the correct response or providing full physical assistance).	Student does not respond or actively resists.

8.4 Interrater Reliability

For the 2006—2007 administration, OPI designed and administered a study to review Interrater Reliability on the Alternate Assessment. A group of five highly qualified administrators observed and scored seven test administrations (for a total of thirty-five students). The scoring was double blind, meaning that the OPI administrators did not communicate their scores to the official test administrator. The two scores from the same administration were analyzed and compared for accuracy. As recommended by the TAC, a highly qualified administrator reviewed and scored thirty evidence templates (in conjunction with their teacher-recording sheets), and those scores were compared to the score given by the actual test administrator. As recommended by the TAC, OPI developed a survey to query the level of training each administrator had received prior to testing. Results from the original Interrater reliability study, the read-behind scoring from the evidence templates, and the survey responses can be found in the paper titled "Examining the Interrater Reliability of Montana's CRT-Alternative" (appendix H).

8.5 Scoring Rules

The instructions and examples illustrate the following rules for scoring:

- Begin with the introductory items and score 4 or 0.
- Use the full scale of 4, 3, 2, 1, and 0 to score the test activity items. Start with level 4 and work systematically through the scaffolding system for every performance indicator as necessary, based on the student's response.
- Allow for appropriate wait time as you scaffold through each level of the scoring rubric.
- Do not repeat questions or directions numerous times.
- Visual, verbal, gestural, and physical cues are allowed in each level except 4.
- Record only one score for each item.
- Score 0 only if the student does not respond or actively resists.
- Halt the administration if the student is showing a pattern of resisting, is becoming fatigued, or is not participating in any way, and resume testing at another time.
- Score every item until the student scores at level 0 for three consecutive items. Stop the administration of the assessment at this point. At the following assessment session, readminister the final three items on which the student scored a 0. If the student receives a

level 0 on three consecutive items again, halt the administration of the assessment and leave the remaining items blank.

8.6 Machine-Scored Items

Once the 2007 test booklets had been logged in, identified with appropriate scannable, preprinted school information sheets, examined for extraneous materials, and batched, they were moved into the scanning area. For all student response booklets (and other forms that required scanning/imaging), this was the last step in the processing loop in which the documents themselves were handled.

At that point, 100 percent of the student response documents and other scannable information necessary to produce the required reports had been captured and converted into an electronic format, including all student identification and demographics, and digital image clips of short-answer and constructed-response student responses. The digital image clip information allowed Measured Progress to replicate student responses on the readers' monitors just as they had appeared on the originals. From that point on, the entire process—data processing, data analysis, and reporting—was accomplished without further reference to the originals.

The first step in that conversion was the removal of the booklet bindings so the individual pages could pass through the scanners one at a time. Once cut, the sheets were put back in their proper boxes and placed in storage until needed for the scanning/imaging process.

Customized scanning programs for all scannables were prepared to selectively read the student response booklets and to electronically format the scanned information according to predetermined requirements. Any information that had been designated time-critical or process-critical was handled first.

In addition to numerous real-time quality control checks, duplex reads, a transport printer that prints a unique identifying number on each sheet of each booklet, and online editing capability, the 5000i scanners offer features that make them compatible with Internet technology.

8.7 Scanning Quality Control

NCS scanners are equipped with many built-in safeguards that prevent data errors. The scanning hardware is continually monitored for conditions that will cause the machine to shut down if

standards are not met. It will display an error message and prevent further scanning until the condition is corrected. The areas monitored include document page and integrity checks, user-designed online edits, and many internal checks of electronic functions.

Before every scanning shift begins, Measured Progress operators perform a daily diagnostic routine. This is yet another step to protect data integrity and one that has been done faithfully for the many years that Measured Progress has been involved in production scanning. In the rare event that the routine detects a photocell that appears to be out of range, Measured Progress will calibrate that machine and perform the test again. If the read is still not up to standard, Measured Progress will call for assistance from our field service engineer.

As a final safeguard, spot checks of scanned files, bubble by bubble and image by image, were routinely made throughout scanning runs. The result of these precautions, from the original layout of the scanning form to the daily vigilance by our operators, was a scan-error rate well below 1 per 1000.

8.8 Electronic Data Files

Once the data had been entered and the scanning logs and other paperwork had been completed, the booklets themselves were put into storage (where they stayed for at least 180 days beyond the close of the fiscal year). When it was determined the files were complete and accurate, those files were duplicated electronically and made available for many other processing options.

Chapter 9—Item Analyses

As noted in Brown (1983), "A test is only as good as the items it contains." A complete evaluation of a test's quality must include an evaluation of each item. Both the *Standards for Educational and Psychological Testing* and the *Code of Fair Testing Practices in Education* include standards for identifying quality items. While the specific statistical criteria identified in these publications were developed primarily for general, not alternate, assessment, the principles—as well as some of the techniques—apply within the alternate assessment framework as well.

Both qualitative and quantitative analyses were conducted to ensure that Montana CRT-Alternate items met these standards. Qualitative analyses are described in earlier sections of this report; this section focuses on the more quantitative evaluations. The statistical evaluations discussed are: difficulty indices, item-test correlations, and differential item functioning (DIF) analyses; note, however, that because of the small sample sizes taking the test, it was not feasible to calculate DIF statistics for the Montana CRT-Alternate. The item analyses presented here are based on the statewide administration of the Montana CRT-Alternate in spring 2007. Table 1 gives the total number of students who participated in each assessment by grade and content area.

Table 1

Number of Students Participating in Each Assessment for Spring 2007		
Grade	**Content Area**	**N**
3	Mathematics	68
	Reading	69
4	Mathematics	90
	Reading	90
5	Mathematics	73
	Reading	71
6	Mathematics	108
	Reading	107
7	Mathematics	73
	Reading	71
8	Mathematics	72
	Reading	72
10	Mathematics	107
	Reading	107

9.1 Difficulty Indices (*p*)

All tasks were evaluated in terms of item difficulty according to standard classical test theory practices. "Difficulty" was defined as the average proportion of points achieved on an item and was measured by obtaining the average score on an item and dividing by the maximum score for the item. Montana CRT-Alternate items are scored polytomously, such that a student can achieve a score of 0, 1, 2, 3, or 4 for an item. By computing the difficulty index as the average proportion of points achieved, the items are placed on a scale that ranges from 0.0 to 1.0. Although this index is traditionally described as a measure of difficulty, it is properly interpreted as an "easiness index" because larger values indicate easier items.

An index of 0.0 indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item. Items that have either a very high or very low difficulty index are considered to be potentially problematic because they are either so difficult that few students get them right or so easy that nearly all students get them right. In either case, such items should be reviewed for appropriateness for inclusion on the assessment. If an assessment was composed entirely of very easy or very hard items, all students would receive nearly the same scores, and the assessment would not be able to differentiate high-ability students from low-ability students. However, it is important to note that the purpose of alternate assessments such as the Montana CRT-Alternate is generally not to differentiate among students, but instead to provide evidence as to how students are progressing relative to performance standards. Therefore, generally accepted criteria regarding item statistics are not applicable to the Montana CRT-Alternate.

9.2 Item-Test Correlations (Item Discrimination)

A desirable feature of an item is that the higher-ability students perform better on the item than lower-ability students. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of an item. Within classical test theory, the item-test correlation is referred to as the item's discrimination because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. The discrimination index used to evaluate Montana CRT-Alternate tasks was the Pearson product-moment correlation. The theoretical range of this statistic is –1.0 to +1.0.

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by other items contributing to the criterion total score. That is, the discrimination index can be thought of as a measure of construct consistency. In light of this interpretation, the selection of an appropriate criterion total score is crucial to the interpretation of the discrimination index. For the Montana CRT-Alternate, the test total score, excluding the item being evaluated, was used as the criterion score.

9.3 Summary of Item Analysis Results

A summary of the item difficulty and item discrimination statistics for each grade/content combination is presented in Table 2. The mean difficulty values shown in Table 2 indicate that, overall, students performed well on the items on the Montana CRT-Alternate. In interpreting these values, it is important to note that item scores lower than 2 are fairly rare on the CRT-Alternate, and a score of 0 is awarded only if the student refuses to respond. These aspects of the item score scale should be considered when evaluating the difficulty values presented in Table 2. In contrast, difficulty values for assessments designed for the general population (i.e., general, rather than alternate, assessments) tend to be in the 0.4 to 0.7 range for the majority of items. Because the nature and purpose of alternate assessments are different from those of general assessments, and because very few guidelines exist as to criteria for interpreting these values for alternate assessments, the values presented in Table 2 should not be interpreted to mean that the students performed better on the CRT-Alternate than the students who took general assessments did on those tests.

Also shown in Table 2 are the mean discrimination values. A couple of factors should be considered when interpreting these values. First, all items on the CRT-Alternate are polytomously scored. In general, polytomous items will tend to have higher discrimination values than dichotomous (e.g., multiple-choice) items because they are less impacted by restriction of range. Second, the item score scale awards points based on the extent to which students require assistance to complete the task. Because students who require assistance with one task are more likely to require assistance on other tasks, discrimination values will be higher for items scored in this way.

As with the item difficulty values, because the nature and use of this assessment are different from those for general assessments, and because very few guidelines exist as to criteria for interpreting these values for alternate assessments, the statistics presented in Table 2 should be interpreted with caution.

Table 2

Item Analysis

Grade	Content Area	Difficulty		Discrimination	
		Mean	SD	Mean	SD
3	Mathematics	0.83	0.10	0.71	0.12
	Reading	0.79	0.12	0.62	0.16
4	Mathematics	0.77	0.11	0.67	0.12
	Reading	0.85	0.11	0.62	0.18
5	Mathematics	0.82	0.10	0.60	0.20
	Reading	0.81	0.12	0.54	0.16
6	Mathematics	0.86	0.08	0.69	0.17
	Reading	0.86	0.09	0.63	0.17
7	Mathematics	0.84	0.12	0.67	0.16
	Reading	0.85	0.11	0.62	0.24
8	Mathematics	0.71	0.13	0.70	0.14
	Reading	0.86	0.09	0.63	0.12
10	Mathematics	0.82	0.11	0.68	0.17
	Reading	0.90	0.07	0.65	0.14

9.4 Differential Item Functioning

Due to very small sample sizes (68 to 108 students in each grade/content combination), it is unreasonable to calculate DIF statistics for the Montana CRT-Alternate. That is, Type I error rates would be unreasonably high and would result in incorrect conclusions regarding the functioning of the items between reference and focal groups. Thus, DIF statistics are not included as part of this technical report.

OPI was responsible for organizing and facilitating committees to review items and reading passages for bias and sensitivity. OPI sent the feedback from the committees to Measured Progress to make the appropriate changes to the items and reading passages.

CHAPTER 10—RELIABILITY

Although an individual item's performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way items function together and complement one another. Tests that function well provide an accurate assessment of the student's level of achievement. Unfortunately, no test can do this perfectly. A variety of factors can contribute to a given student's score being either higher or lower than his or her true achievement. Collectively, these extraneous factors that impact a student's score are referred to as measurement error. Any assessment includes some amount of measurement error; that is, no measurement can be perfectly accurate. This is true of academic assessments—no assessment can measure students perfectly accurately; some students will receive scores that underestimate their true achievement, and other students will receive scores that overestimate their true achievement. When tests have a high amount of measurement error, student scores are very unstable. Students with high achievement may get low scores, or vice versa. Consequently, one cannot reliably tell a student's true level of achievement with such a test. Assessments that have less measurement error (i.e., errors made are small on average and student scores on such a test will consistently represent their achievement) are described as reliable.

There are a number of ways to estimate an assessment's reliability. One possible approach is to give the same test to the same students at two different points in time. If students receive the same scores on each test, then the extraneous factors affecting performance are small and the test is reliable (this is referred to as test-retest reliability). A potential problem with this approach is that students may remember items from the first administration or may have gained (or lost) knowledge or skills in the interim between the two administrations. A solution to the "remembering items" problem is to give a different, but parallel test, at the second administration. If student scores on each test correlate highly, the test is considered reliable (this is known as alternate-forms reliability because an alternate form of the test is used in each administration). This approach, however, does not address the problem that students may have gained (or lost) knowledge or skills in the interim between the two administrations. In addition, the practical challenges of developing and administering parallel forms generally preclude the use of parallel forms reliability indices. One way to address these problems is to split the test in half and then correlate students' scores on the two half-tests; this in effect treats each half-test as a complete test. By doing this, the problems associated with an intervening time interval, and of creating and administering two parallel forms of the test, are alleviated. This is known as a split-half estimate

of reliability. If the two half-test scores correlate highly, items on the two half-tests must be measuring very similar knowledge or skills. This is evidence that the items complement one another and function well as a group. This also suggests that measurement error will be minimal.

The split-half method requires a judgment regarding the selection of which items contribute to which half-test score. This decision may have an impact on the resulting correlation; different splits will give different estimates of reliability. Cronbach (1951) provided a statistic, α, that avoids this concern about the split-half method. Cronbach's α gives an estimate of the average of all possible splits for a given test. Cronbach's α is often referred to as a measure of internal consistency because it provides a measure of how well all the items in the test measure one single underlying ability. Cronbach's α is computed using the following formula:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma^2_{(Y_i)}}{\sigma^2_x}\right]$$

where i indexes the item,
n is the total number of items,
$\sigma^2(Y_i)$ represents individual item variance, and
σ^2_x represents the total test variance.

10.1 Reliability Results

Table 3 presents Cronbach's α coefficient for each subject area (reading and mathematics) and each grade level. The values in Table 3 are all greater than or equal to 0.89, indicating that these tests have a high level of reliability. Note, however, that these high values do not necessarily indicate that the CRT-Alternate is "better" than general assessments, which tend to have reliabilities ranging from around 0.80 to around 0.95. There are several factors that may contribute to these high values. First, because the CRT-Alternate is individually administered, the reliability values are likely to be inflated due to administrator effects. In other words, the task scores awarded by the administrator may be

influenced by his or her overall sense of the student's level of ability or proficiency, which may result in task scores that are more homogeneous than they would be if they were based strictly on the student's performance on each task. Second, the reliabilities are artificially inflated due to the fact that tasks are "bundled" together within activities. Items that are bundled together will be more highly correlated, which will increase test reliability. Finally, the use of level of assistance required in the item scoring guide (as described above) will also increase item intercorrelations and overall test reliability.

Table 3

Reliability Analysis

Grade	Content Area	Reliability
3	Mathematics	0.95
	Reading	0.94
4	Mathematics	0.95
	Reading	0.92
5	Mathematics	0.94
	Reading	0.89
6	Mathematics	0.95
	Reading	0.94
7	Mathematics	0.94
	Reading	0.93
8	Mathematics	0.96
	Reading	0.90
10	Mathematics	0.95
	Reading	0.92

10.2 Reliability of Performance-Level Categorization

All test scores contain measurement error; thus, classifications based on test scores are also subject to measurement error. After the performance levels were specified and students were classified into those levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications. For the Montana CRT-Alternate, students are classified into one of four performance levels: *Novice* (N), *Nearing Proficiency* (NP), *Proficient* (P), or *Advanced* (A). This section of the report explains the methodologies used to assess the reliability of classification decisions, and results are given.

Accuracy

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated because errorless test scores do not exist.

Consistency

Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete, parallel forms of the test are given to the same group of students. This is usually impractical, especially on lengthy tests. To overcome this issue, techniques have been developed to estimate both accuracy and consistency of classification decisions based on a single administration of a test. The technique developed by Livingston and Lewis (1995) was used for the Montana CRT-Alternate because it is a flexible approach that is appropriate for tests that are composed entirely of polytomous items.

Calculating Accuracy

All of the accuracy and consistency estimation techniques described below make use of the concept of "true scores" in the sense of classical test theory. A true score is the score that would be obtained on a test that had no measurement error. It is a theoretical concept that cannot be observed, although it can be estimated. In the Livingston and Lewis method, the estimated true score distribution is used to estimate the proportion of students in each "true" performance level. After various technical adjustments (which are described in Livingston and Lewis, 1995), a 4×4 contingency table was created for each content area and grade level. The $[i,j]$ entry of an accuracy table represents the estimated proportion of students whose true score fell into performance level i and whose observed score fell into performance level j on the Montana CRT-Alternate. Overall accuracy, which is the proportion of students whose true and observed performance levels match one another, is the sum of the numbers on the diagonal of the accuracy table.

Calculating Consistency

To estimate consistency, the true scores are used to estimate the joint distribution of classifications on two independent, parallel test forms. After statistical adjustments (see Livingston and Lewis, 1995), a new 4 × 4 contingency table was created for each content area and grade level that shows the proportion of students who would be classified into each performance level by the two (hypothetical) parallel test forms. That is, the [*i*,*j*] entry of a consistency table represents the estimated proportion of students whose observed score on the first form would fall into performance level *i* and whose observed score on the second form would fall into performance level *j*. Overall consistency, which is the proportion of students classified into exactly the same performance level by the two forms of the test, is the sum of the numbers on the diagonal of this new contingency table.

Kappa

Another way to measure consistency is to use Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. Cohen's κ can be used to evaluate the classification consistency of a test from two parallel forms of the test. The two forms in this case were the hypothetical parallel forms used by the Livingston and Lewis method. Because κ is corrected for chance, the values of κ are lower than other consistency estimates.

Results of Accuracy, Consistency, and Kappa Analyses

Summaries of the accuracy and consistency analyses are provided in Tables 4 through 17. The first section of each table shows the overall accuracy and consistency indices, as well as κ. The overall index, as described above, is the sum of the diagonal elements of the appropriate contingency table, and κ, as described above, is a version of the overall consistency value that has been corrected for chance. Note that, as expected, the values of κ reported in Tables 4 through 17 are lower than the overall consistency estimates.

The second section of each table shows accuracy and consistency values conditional upon performance level. In each case, the denominator is the number of students who are associated with a given performance level. For example, the conditional accuracy value is 0.8031 for the *Proficient* level for grade 4 math. This figure indicates that among the students whose true scores placed them in the

Proficient level, 80.31% of them would be expected to be placed in *Proficient* if they were categorized according to their observed scores. The corresponding consistency value of 0.7734 indicates that 77.34% of students with observed scores in the *Proficient* performance level would be expected to score in *Proficient* again if a second, parallel test form were used.

For certain tests, concern may be greatest regarding decisions made about a particular threshold. For example, for purposes of accountability, there is generally greatest interest in distinguishing between students who are *Proficient* or *Advanced* and those who have not yet reached the *Proficient* threshold. The third section of the summary tables shows information at each of the cut points. These values indicate the accuracy and consistency of the dichotomous decisions, either above or below the associated cut point. In addition, the false-positive and false-negative accuracy rates are also provided. These values are estimates of the proportion of students who were categorized above the cut when their true score would place them below the cut (false positive), and vice versa.

10.3 Examining the Interrater reliability

During the peer review phase of NCLB, a request to study the interrater reliability of the CRT-Alternate was made by the U.S. Department of Education. The resulting study was designed to produce a "preponderance of evidence" supporting the overall integrity as well as the interrater reliability of the CRT-Alt. The resulting paper of the study, produced by Gail McGregor of the University of Montana—Missoula, is titled "Examining the Interrater Reliability of Montana's CRT-Alternative" and has been attached to this technical manual as Appendix H. The study reviewed the following areas:

- Evidence-Base for Practices used in Test Design
- Accessibility of Training for Test Administrators
- Test Administrator Knowledge and Understanding of Testing Procedures
- Fidelity of Test Administration
- Level of Agreement: Item Scoring

With regard to these five areas, it was concluded that the assessment had a high degree of integrity but could be improved in some areas. The study showed the test design was appropriate for the students being assessed. The training methods seem to be sufficient reaching the test

administrators, although the survey of teachers was often times redundant causing unnecessary work. The self-check tools appeared to be beneficial and should be continued in the future. The direct observation of a sample of test administrators and the evidence template review were both positive, though with greater resources, a larger sample size would produce results with a higher confidence interval.

TABLE 4
ACCURACY AND CONSISTENCY — GRADE 3 MATH

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.7623		0.7170		0.5977
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.8935		0.8785
	Nearing Proficiency		0.5068		0.4095
	Proficient		0.4846		0.4110
	Advanced		0.9199		0.8187
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9442	0.0349	0.0209	0.9239
	NP : P	0.9291	0.0489	0.0220	0.9060
	P : A	0.8802	0.0939	0.0259	0.8581

TABLE 5
ACCURACY AND CONSISTENCY — GRADE 4 MATH

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.8176		0.7599		0.6629
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.8807		0.8464
	Nearing Proficiency		0.6141		0.5040
	Proficient		0.8031		0.7734
	Advanced		0.9211		0.8045
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9541	0.0254	0.0205	0.9360
	NP : P	0.9406	0.0351	0.0244	0.9176
	P : A	0.9221	0.0614	0.0165	0.9005

TABLE 6
ACCURACY AND CONSISTENCY — GRADE 5 MATH

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.7669		0.7180		0.6047
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.8858		0.8653
	Nearing Proficiency		0.3900		0.2995
	Proficient		0.6483		0.6042
	Advanced		0.9150		0.7905
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9404	0.0361	0.0234	0.9182
	NP : P	0.9314	0.0442	0.0244	0.9069
	P : A	0.8840	0.0945	0.0215	0.8634

TABLE 7
ACCURACY AND CONSISTENCY — GRADE 6 MATH

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.7816		0.7352		0.6102
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.8818		0.8567
	Nearing Proficiency		0.7087		0.6412
	Proficient		0.4360		0.3532
	Advanced		0.9261		0.8467
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9625	0.0221	0.0154	0.9483
	NP : P	0.9340	0.0456	0.0203	0.9133
	P : A	0.8778	0.0912	0.0309	0.8532

TABLE 8
ACCURACY AND CONSISTENCY— GRADE 7 MATH

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.8137		0.7623		0.6513
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.8498		0.8056
	Nearing Proficiency		0.6850		0.5916
	Proficient		0.7069		0.6739
	Advanced		0.9479		0.8640
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9665	0.0185	0.0150	0.9535
	NP : P	0.9469	0.0324	0.0207	0.9272
	P : A	0.9002	0.0798	0.0200	0.8793

TABLE 9
ACCURACY AND CONSISTENCY— GRADE 8 MATH

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.8344		0.7738		0.6922
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.9010		0.8729
	Nearing Proficiency		0.7058		0.6079
	Proficient		0.7130		0.6229
	Advanced		0.9369		0.8797
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9527	0.0263	0.0210	0.9339
	NP : P	0.9422	0.0344	0.0235	0.9197
	P : A	0.9394	0.0401	0.0206	0.9169

TABLE 10
ACCURACY AND CONSISTENCY— GRADE 10 MATH

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.8149		0.7602		0.6599
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.8854		0.8581
	Nearing Proficiency		0.6900		0.5993
	Proficient		0.6009		0.5199
	Advanced		0.9484		0.8787
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9568	0.0250	0.0182	0.9401
	NP : P	0.9390	0.0390	0.0219	0.9169
	P : A	0.9178	0.0620	0.0201	0.8953

TABLE 11
ACCURACY AND CONSISTENCY — GRADE 3 READING

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.7952		0.7315		0.6313
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.8483		0.8016
	Nearing Proficiency		0.7608		0.6929
	Proficient		0.6826		0.6231
	Advanced		0.9251		0.8206
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9608	0.0213	0.0178	0.9456
	NP : P	0.9307	0.0434	0.0259	0.9052
	P : A	0.9034	0.0752	0.0214	0.8783

TABLE 12
ACCURACY AND CONSISTENCY — GRADE 4 READING

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.7738		0.7134		0.5801
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.8323		0.7833
	Nearing Proficiency		0.6169		0.5177
	Proficient		0.6165		0.5517
	Advanced		0.9359		0.8452
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9573	0.0236	0.0191	0.9410
	NP : P	0.9294	0.0441	0.0265	0.9041
	P : A	0.8854	0.0891	0.0255	0.8573

TABLE 13
ACCURACY AND CONSISTENCY — GRADE 5 READING

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.7554		0.6911		0.5374
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.7894		0.7155
	Nearing Proficiency		0.6505		0.5615
	Proficient		0.4962		0.4084
	Advanced		0.9274		0.8478
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9525	0.0248	0.0227	0.9342
	NP : P	0.9106	0.0561	0.0334	0.8791
	P : A	0.8852	0.0828	0.0320	0.8504

TABLE 14
ACCURACY AND CONSISTENCY — GRADE 6 READING

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.8399		0.7896		0.6637
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.8415		0.7896
	Nearing Proficiency		0.6812		0.5851
	Proficient		0.7100		0.6565
	Advanced		0.9547		0.8981
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9726	0.0146	0.0127	0.9619
	NP : P	0.9526	0.0281	0.0194	0.9347
	P : A	0.9145	0.0630	0.0225	0.8906

TABLE 15
ACCURACY AND CONSISTENCY — GRADE 7 READING

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.8418		0.7904		0.6448
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.8083		0.7429
	Nearing Proficiency		0.6699		0.5720
	Proficient		0.6957		0.6278
	Advanced		0.9538		0.9049
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9738	0.0139	0.0124	0.9636
	NP : P	0.9520	0.0281	0.0199	0.9340
	P : A	0.9157	0.0590	0.0253	0.8895

TABLE 16
ACCURACY AND CONSISTENCY — GRADE 8 READING

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.7560		0.7186		0.5600
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.8226		0.7882
	Nearing Proficiency		0.4655		0.3766
	Proficient		0.4287		0.3665
	Advanced		0.9537		0.8711
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9399	0.0364	0.0237	0.9184
	NP : P	0.9181	0.0562	0.0257	0.8925
	P : A	0.8827	0.0961	0.0213	0.8622

TABLE 17
ACCURACY AND CONSISTENCY — GRADE 10 READING

Accuracy and Consistency of Classification Indices					
Overall Indices	**Accuracy**		**Consistency**		**Kappa (κ)**
	0.8430		0.8125		0.6405
Indices Conditional on Level			**Accuracy**		**Consistency**
	Novice		0.8439		0.8145
	Nearing Proficiency		0.4716		0.3760
	Proficient		0.5396		0.4763
	Advanced		0.9732		0.9282
Indices for Dichotomous Decisions Around Cut Points		**Accuracy**			**Consistency**
		Accuracy	***False Positives***	***False Negatives***	
	N : NP	0.9643	0.0217	0.0140	0.9513
	NP : P	0.9540	0.0302	0.0158	0.9384
	P : A	0.9199	0.0636	0.0165	0.9050

Chapter 11—Scaling

11.1 Translating Raw Scores to Scaled Scores and Performance Levels

Montana CRT-Alternate scores in each content area are reported on a scale that ranges from 200 to 300. Scaled scores supplement the Montana CRT-Alternate performance-level results by providing information about the position of a student's results within a performance level. School- and district-level scaled scores are calculated by computing the average of student-level scaled scores. Students' raw scores, or total number of points, on the Montana CRT-Alternate tests are translated to scaled scores using a data analysis process called scaling. Scaling simply converts raw points from one scale to another. In the same way that the same temperature can be expressed on either the Fahrenheit or Celsius scales and the same distance can be expressed either in miles or kilometers, student scores on the Montana CRT-Alternate tests can be expressed as raw scores or scaled scores.

It is important to note that converting from raw scores to scaled scores does not change the students' performance-level classifications. Given the relative simplicity of raw scores, it is fair to ask why scaled scores are used in Montana CRT-Alternate reports instead of raw scores. Foremost, scaled scores offer the advantage of simplifying the reporting of results across content areas, grade levels, and subsequent years. Because the standard setting process typically results in different cut scores across content areas on a raw score basis, it is useful to transform these raw cut scores to a scale that is more easily interpretable and consistent. For the Montana CRT-Alternate, a score of 225 is the cut score between the *Novice* and *Nearing Proficiency* performance levels. This is true regardless of which content area, grade, or year one may be concerned with. If one were to use raw scores, the raw cut score between *Novice* and *Nearing Proficiency* may be, for example, 35 in mathematics at grade 8, but 33 in mathematics at grade 10, or 36 in reading at grade 8. Using scaled scores greatly simplifies the task of understanding how a student performed.

Raw score cut points for the Montana CRT-Alternate were established via standard setting in July 2006. Details of the standard setting are documented in the standard setting report, which is included in appendix C. Once the 2006 raw score cut points were determined, the next step was to calculate the transformation coefficients that would be used to place students' raw scores onto the score scale used for reporting. As previously stated, student scores on the Montana CRT-Alternate are reported in integer values from 200 to 300, with three scores representing cut scores on each assessment. Two of the three cut points (*Novice*/*Nearing Proficiency* and *Nearing*

Proficiency/*Proficient*) were pre-set at 225 and 250, respectively; the third cut point, between *Proficient* and *Advanced*, was allowed to vary across tests, depending on where the raw score cuts were placed. Allowing the upper cut to float results in a single conversion equation for each test, which simplifies interpretation of scaled scores and their summary statistics. Table 18 presents the scaled score range for each performance level in each grade/content area combination.

Table 18

Scaled Score Ranges

Grade	Content Area	Scaled Score Range for each Performance Level			
		Novice	***Nearing Proficiency***	***Proficient***	***Advanced***
3	Mathematics	200–224	225–249	250–268	269–300
	Reading	200–224	225–249	250–264	265–300
4	Mathematics	200–224	225–249	250–294	295–300
	Reading	200–224	225–249	250–270	271–300
5	Mathematics	200–224	225–249	250–296	297–300
	Reading	200–224	225–249	250–262	263–300
6	Mathematics	200–224	225–249	250–257	258–300
	Reading	200–224	225–249	250–274	275–300
7	Mathematics	200–224	225–249	250–274	275–300
	Reading	200–224	225–249	250–276	277–300
8	Mathematics	200–224	225–249	250–272	273–300
	Reading	200–224	225–249	250–268	269–300
10	Mathematics	200--224	225--249	250--264	265--300
	Reading	200--224	225--249	250--277	278--300

The scaled scores are obtained by a simple linear transformation of the raw scores using the values of 225 and 250 on the scaled score metric and the associated 2006 raw score cut points to define the transformation. The scaling coefficients were calculated using the following formulae:

$$b = 225 - m(x_1)$$

$$b = 250 - m(x_2)$$

$$m = \frac{225 - 250}{x_1 - x_2}$$

where m is the slope of the line providing the relationship between the raw and scaled scores, b is the intercept, x_1 is the cut score on the raw score metric for the *Novice*/*Nearing Proficiency* cut, and x_2 is

the cut score on the raw score metric for the *Nearing Proficiency*/*Proficient* cut. The raw score cut points (x_1 and x_2) were determined in the standard setting meeting held in July 2006 (see the 2006 standard setting report for details). Scaled scores were then calculated using the following linear transformation:

$$ss = m(x) + b$$

where x represents a student's raw score. The values obtained using this formula were rounded to the nearest integer and truncated, as necessary, such that no student received a score below 200 or higher than 300.

Chapter 12—Reporting

The CRT-Alternate assessments were designed to measure student performance against Montana's Content Standards and Expanded Benchmarks. Consistent with this purpose, results on the CRT-Alternate were reported in terms of performance levels that describe student performance in relation to these established state standards. There are four performance levels: *Advanced*, *Proficient*, *Nearing Proficiency*, and *Novice*. (CRT-Alternate performance level descriptors, scaled score ranges, and raw scores are described in greater detail in appendix D.) Students receive a separate performance-level classification (based on total scaled score) in each content area.

School- and system-level results are reported as the number and percentage of students attaining each performance level at each grade level tested. Disaggregations of students are also reported at the school and system levels. The CRT-Alternate reports are

- Student Reports;
- Class Roster & Item-Level Reports;
- School Summary Reports; and
- System Summary Reports.

"Decision Rules" were formulated in late spring 2007 by OPI and Measured Progress to identify students, during the reporting process, to be excluded from school- and system-level reports. A copy of these decision rules is included in this report as appendix G.

State summary results were provided to OPI on confidential CDs and via a secure Web site. The report formats are included in appendix F. These reports were shipped to system test coordinators in June 2007 for distribution to schools within their respective systems/districts. System test coordinators and teachers were also provided with copies of the *Guide to Interpreting the 2007 Criterion-Referenced Test and CRT-Alternate Assessment Reports* to assist them in understanding the connection between the assessment and the classroom. The guide provides information about the assessment and the use of assessment results.

CHAPTER 13—VALIDITY SUMMARY

The purpose of this manual is to describe several technical aspects of the CRT-Alternate in an effort to contribute to the accumulation of validity evidence to support CRT-Alternate score interpretations. Because it is the interpretations of test scores that are evaluated for validity, not the test itself, this manual presents documentation to substantiate intended interpretations (AERA, 1999). Each of the chapters in this manual contributes important information to the validity argument by addressing one or more of the following aspects of the CRT-Alternate: test development, test alignment, test administration, scoring, item analyses, reliability, scaling, performance levels, and reporting.

The CRT-Alternate assessments are based on, and aligned to, Montana's Content Standards and Expanded Benchmarks in reading and mathematics. Intended inferences from the CRT-Alternate results are about student achievement on Montana's reading and mathematics Content Standards and Expanded Benchmarks, and these achievement inferences are meant to be useful for program and instructional improvement and as a component of school accountability.

The *Standards for Educational and Psychological Testing* (1999) provides a framework for describing sources of evidence that should be considered when constructing a validity argument. These sources include evidence based on the following five general areas: test content, response processes, internal structure, relationship to other variables, and consequences of testing. Although each of these sources may speak to a different *aspect* of validity, they are not distinct *types* of validity. Instead, each contributes to a body of evidence about the comprehensive validity of score interpretations.

A measure of test content validity is to determine how well the assessment tasks represent the curriculum and standards for each subject and grade level. This is informed by the item development process, including how the test blueprints and test items align to the curriculum and standards. Viewed through this lens provided by the content standards, evidence based on test content was extensively described in chapters 2 through 7. Item alignment with Montana Content Standards; item bias, sensitivity, and content appropriateness review processes; adherence to the test blueprint; use of standardized administration procedures; and appropriate test administration training are all components of validity evidence based on test content. As discussed earlier, all CRT-Alternate test questions are aligned by Montana educators to specific Montana Content Standards and undergo several rounds of review for content fidelity and appropriateness. Finally, tests are administered

according to state-mandated standardized procedures, and all test administrators are required to review the training CD.

The scoring information in chapter 8 describes the steps taken to train the teachers administering the assessment on scoring procedures, as well as quality control procedures related to scanning. In order to obtain additional validity evidence, it would be helpful to conduct a study in which a percentage of teachers administering the assessment would be videotaped to confirm validity of administration and scoring.

Evidence based on internal structure is presented in the discussions of item analyses and reliability in chapters 9 and 10. Technical characteristics of the internal structure of the assessments are presented in terms of classical item statistics (item difficulty, item-test correlation) and reliability coefficients. In general, item difficulty and discrimination indices were in acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that most items were assessing consistent constructs, and students who performed well on individual items tended to perform well overall.

To further support the validity argument, additional studies to provide evidence regarding the relationship of CRT-Alternate results to other variables include the extent to which scores from the CRT-Alternate assessments converge with other measures of similar constructs, and the extent to which they diverge from measures of different constructs. Relationships among measures of the same or similar constructs can sharpen the meaning of scores and appropriate interpretations by refining the definition of the construct.

The evidence presented in this manual supports inferences of student achievement on the content represented in the Montana Content Standards for reading and mathematics for the purposes of program and instructional improvement and as a component of school accountability.

SECTION IV: REFERENCES

American Educational Research Association (AERA), American Psychological Association, and National Council on Measurement in Education. 1999. *Standards for educational and psychological testing*. Washington, DC: AERA.

Brown, F. G. 1983. *Principles of educational and psychological testing.* 3rd ed. Fort Worth, TX: Holt, Rinehart, and Winston.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37–46.

Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334.

Joint Committee on Testing Practices. 2004. *Code of fair testing practices in education.* Washington, DC: American Psychological Association. Available for download at http://www.apa.org/science/fairtestcode.html.

Livingston, S. A., and Lewis, C. 1995. Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement* 32: 179–197.